

Zeroth-order Asynchronous Doubly Stochastic Algorithm with Variance Reduction

Bin Gu
Zhouyuan Huo
Heng Huang

*Department of Computer Science and Engineering
University of Texas at Arlington*

JSGUBIN@GMAIL.COM
ZHOUYUAN.HUO@MAVS.UTA.EDU
HENG@UTA.EDU

Abstract

Zeroth-order (derivative-free) optimization attracts a lot of attention in machine learning, because explicit gradient calculations may be computationally expensive or infeasible. To handle large scale problems both in volume and dimension, recently asynchronous doubly stochastic zeroth-order algorithms were proposed. The convergence rate of existing asynchronous doubly stochastic zeroth order algorithms is $O(\frac{1}{\sqrt{T}})$ (also for the sequential stochastic zeroth-order optimization algorithms). In this paper, we focus on the finite sums of smooth but not necessarily convex functions, and propose an asynchronous doubly stochastic zeroth-order optimization algorithm using the accelerated technology of variance reduction (AsyDSZOVR). Rigorous theoretical analysis show that the convergence rate can be improved from $O(\frac{1}{\sqrt{T}})$ the best result of existing algorithms to $O(\frac{1}{T})$. Also our theoretical results is an improvement to the ones of the sequential stochastic zeroth-order optimization algorithms.

Keywords: stochastic optimization, zeroth-order, parallel computing, lock-free

1. Introduction

Zeroth-order (derivative-free) optimization attracts a lot of attention in machine learning, because explicit gradient calculations may be computationally expensive or infeasible. As we know, for a lot of machine learning optimization problems, such as graphical model inference (Wainwright and Jordan, 2008), structured-prediction (Taskar et al., 2005), and so on, it is difficult to give the explicit derivatives for the objective functions. For some black box learning model, such as black box neural networks (Lian et al., 2016), it is infeasible to give the explicit derivatives. Also, for bandit problems (Bubeck and Cesa-Bianchi, 2012), such as advertisement selection for search engines, it is infeasible to give the explicit derivatives of the objective functions because only observations of function values are available. Since zeroth-order methods estimate gradient based on only two point observations, it is the best and only choice of the optimization for above scenarios.

Because the era of big data has arrived, asynchronous parallel algorithms for stochastic optimization have received huge successes in theory and practice recently. Most of these asynchronous parallel stochastic algorithms are built on the first-order derivative or second-order information (e.g. (approximate) Hessian matrix) of the objective function. For

example, Hogwild! (Recht et al., 2011) (the first lock-free asynchronous parallel stochastic gradient descent (SGD) algorithm) uses the first-order derivative to update the solution for smooth convex functions. The other variants of asynchronous parallel SGD algorithm (Mania et al., 2015; Lian et al., 2015; Huo and Huang, 2016; Zhao and Li, 2016) also use the first-order derivative to update the solution for smooth convex or nonconvex functions. For a composite of a smooth (possibly non-convex) function and a non-smooth convex function, the first-order derivative is embedded in the proximal operator (Razaviyayn et al., 2014; Liu and Wright, 2015; You et al., 2016). Also, second-order information (e.g. (approximate) Hessian matrix) (Byrd et al., 2016) can be used to accelerate the optimization.

As the reasons mentioned previously, designing asynchronous stochastic zeroth order algorithms is important and urgent. As far as we know, the only work of asynchronous stochastic zeroth order algorithm (AsySZO) is (Lian et al., 2016). They prove the convergence rate $O(\frac{1}{T} + \frac{1}{\sqrt{T}})$. To the best of our knowledge, the convergence rates of existing sequential stochastic zeroth order algorithms (Nesterov and Spokoiny, 2011; Jamieson et al., 2012; Duchi et al., 2012; Agarwal et al., 2011) are $O(\frac{1}{T} + \frac{1}{\sqrt{T}})$ or $O(\frac{1}{\sqrt{T}})$. Basically, the convergence rates of these algorithms can be viewed as $O(\frac{1}{\sqrt{T}})$ because the term $\frac{1}{\sqrt{T}}$ dominates $\frac{1}{T} + \frac{1}{\sqrt{T}}$. Motivated by improving the convergence rate of SGD from $O(\frac{1}{\sqrt{T}})$ to $O(\frac{1}{T})$, it is highly desirable to design an accelerated asynchronous stochastic zeroth order algorithm with the convergence rate $O(\frac{1}{T})$.

In this paper, we focus on the finite sums of smooth but not necessarily convex functions as follows.

$$\min_{x \in \mathbb{R}^N} f(x) = \frac{1}{l} \sum_{i=1}^l f_i(x) \quad (1)$$

where $f_i : \mathbb{R}^N \mapsto \mathbb{R}$ is a smooth, possibly non-convex function. The formulation (1) covers an extensive number of machine learning problems, for example, logistic regression (Freedman, 2009), ridge regression (Shen et al., 2013), least squares SVM (Suykens and Vandewalle, 1999) and so on.

In this paper, we propose an asynchronous doubly stochastic zeroth-order optimization algorithm using the accelerated technology of variance reduction (AsyDSZOVR). Our AsyDSZOVR randomly select a set of samples and a set of features simultaneously to handle large scale problems both in volume and dimension. Rigorous theoretical analysis show that the convergence rate can be improved from $O(\frac{1}{\sqrt{T}})$ the best result of existing algorithms to $O(\frac{1}{T})$. Also our theoretical results is an improvement to the ones of the sequential stochastic zeroth-order optimization algorithms.

We organize the rest of the paper as follows. In section 2, we propose our AsySBCDVR algorithm. In Section 3, we prove the convergence rate for AsySBCDVR. Finally, we give some concluding remarks in Section 4.

2. Algorithms

In this section, we propose our AsyDSZOVR. In this paper, we focus on the parallel environment with shared memory, such as multi-core processors and GPU-accelerators, without any

lock. Because the parallel computing pattern in the parallel environment with distributed memory can be equivalent to the one in the parallel environment with shared memory having reading and writing locks, our AsyDSZOVR can also work in the parallel environment with distributed memory.

The basic parallel computing pattern includes three steps, i.e., read, compute, update. Specifically, if the parallel computing is asynchronous, all cores repeat the three steps independently and concurrently without any lock. We give a more detailed descriptions of the three steps as following.

1. **Read:** Read the vector x from the shared memory to the local memory without reading lock.
2. **Compute:** Randomly choose a component function f_i or a mini-batch \mathcal{B} of the component functions, and a set of coordinates J , and locally compute an unbiased (approximate) gradient.
3. **Update:** Update the set of coordinates J of the vector x in the shared memory, based on the unbiased (approximate) gradient without writing lock.

To highlight the differences of AsySZO and our proposed AsyDSZOVR, we first give brief review of AsySZO, and present our AsyDSZOVR based on the above framework of parallel computing. We also summarize the differences of of AsySZO and AsyDSZOVR in Table 1.

Table 1: Comparisons of AsySZO and AsyDSZOVR.

Algorithm	Accelerated	Step size	Mini-batch	$\hat{x}_t - x_t$ or $\hat{x}_t^{s+1} - x_t^{s+1}$	Rate
AsySZO	No	Dynamic vanishing	No	$\gamma \sum_{t' \in K(t)} G_{J(t')}(\hat{x}_t; f_i)$	$O\left(\frac{1}{\sqrt{T}}\right)$
AsyDSZOVR	Yes	Constant	Yes	$\gamma \sum_{t' \in K(t)} B_{t'}^{s+1} \hat{v}_{J(t')}^{s+1}$	$O\left(\frac{1}{T}\right)$

2.1 Brief Review of AsySZO

Actually, the existing asynchronous stochastic zeroth order algorithm (i.e., AsySZO) proposed by (Lian et al., 2016) strictly follows the three steps. Specifically, the unbiased (approximate) gradient in the ‘**Compute**’ step is computed based on a randomly choosed component function f_i as

$$G_J(x; f_i) = \sum_{j \in J} \frac{N}{2Y\mu_j} (f_i(x + \mu_j e_j) - f_i(x - \mu_j e_j)) e_j \quad (2)$$

where μ_j is the approximate parameter for the j -th coordinate, and e_j is the zero vector in \mathbb{R}^N except that the coordinates indexed by j equal to 1. Thus, the updating rule in the ‘**Update**’ step is $(x_{t+1}^{s+1})_J \leftarrow ((x_t^{s+1}) - \gamma G_J(\hat{x}_t^{s+1}; f_i))_J$, where γ is the step size. The pseudocode of AsySZO can be found in Algorithm 1.

Because AsySZO does not use the reading and writing locks, the vector \hat{x}_t^{s+1} read into the local memory may be inconsistent to the vector x_t^{s+1} in the shared memory, which means

that some components of \hat{x}_t^{s+1} are same with the ones in x_t^{s+1} , but others are different to the ones in x_t^{s+1} . In (Lian et al., 2016), they present x_t^{s+1} as following.

$$x_t = \hat{x}_t - \gamma \sum_{t' \in K(t)} G_{J(t')}(\hat{x}_t; f_i) \quad (3)$$

where $K(t)$ is a set of iterations. As mentioned in (Mania et al., 2015; Zhao and Li, 2016), this representation could not formulate the conflicts of two writing operations. For AsyDSZOVR, we will give a more reasonable representation of x_t^{s+1} .

Algorithm 1 Asynchronous Stochastic Zeroth-order Optimization (AsySZO)

Input: γ , S , and m .

Output: x^S .

- 1: Initialize $x^0 \in \mathbb{R}^d$, p threads.
 - 2: *For each thread*, do:
 - 3: **for** $t = 0, 1, 2, m - 1$ **do**
 - 4: Randomly select a component function f_i from $\{1, \dots, l\}$ with equal probability.
 - 5: Randomly choose a set of coordinates $J(t)$ from $\{1, \dots, n\}$ with equal probability.
 - 6: $(x_{t+1}^{s+1})_{J(t)} \leftarrow ((x_t^{s+1}) - \gamma G_{J(t)}(\hat{x}_t^{s+1}; f_i))_{J(t)}$.
 - 7: $(x_{t+1}^{s+1})_{\setminus J(t)} \leftarrow (x_t^{s+1})_{\setminus J(t)}$.
 - 8: **end for**
-

2.2 AsyDSZOVR

Although $G_J(x; f_i)$ is an unbiased estimate of $G_J(x; f)$, it would have a large variance because it is computed based on one sample. Similar with (Huo and Huang, 2016; Zhao and Li, 2016), we use the variance reduction to accelerate AsySZO. Thus, AsyDSZOVR has two-layer loops. The outer layer is to parallelly compute the full approximate gradient $G_J(x^s; f) = \frac{1}{l} \sum_{i=1}^l G_J(x^s; f_i)$, where the superscript s denotes the s -th outer loop. The inner layer is to parallelly and repeatedly update the vector x in the shared memory, which also strictly follows the three steps as mentioned previously. Specifically, all cores repeat the following steps independently and concurrently without any lock:

1. **Read:** Read the vector x from the shared memory to the local memory without reading lock. We use \hat{x}_t^{s+1} to denote its value, where the subscript t denotes the t -th inner loop.
2. **Compute:** Randomly choose a mini-batch $\mathcal{B}(t)$ of the component functions, and a set of coordinates $J(t)$ from $\{1, \dots, N\}$, and locally compute $\hat{v}_{J(t)}^{s+1} = \frac{1}{|\mathcal{B}(t)|} \sum_{i \in \mathcal{B}(t)} G_{J(t)}(\hat{x}_t^{s+1}; f_i) - \frac{1}{|\mathcal{B}(t)|} \sum_{i \in \mathcal{B}(t)} G_{J(t)}(\tilde{x}^s; f_i) + G_{J(t)}(\tilde{x}^s; f)$.
3. **Update:** Update the set of coordinates $J(t)$ of the vector x in the shared memory as $(x_{t+1}^{s+1})_{J(t)} \leftarrow ((x_t^{s+1}) - \gamma \hat{v}_{J(t)}^{s+1})_{J(t)}$ without writing lock.

The detailed description of AsyDSZOVR is presented in Algorithm 2. Note that $\hat{v}_{J(t)}^{s+1}$ computed locally is an approximation of $G_J(\hat{x}_t^{s+1}; f)$, and the expectation of $\hat{v}_{J(t)}^{s+1}$ on $\mathcal{B}(t)$

is equal to $G_J(\hat{x}_t^{s+1}; f)$ as shown below.

$$\begin{aligned}
\mathbb{E}_{\mathcal{B}(t)} \hat{v}_{J(t)}^{s+1} &= \mathbb{E}_{\mathcal{B}(t)} \left(\frac{1}{|\mathcal{B}(t)|} \sum_{i \in \mathcal{B}(t)} G_{J(t)}(\hat{x}_t^{s+1}; f_i) - \frac{1}{|\mathcal{B}(t)|} \sum_{i \in \mathcal{B}(t)} G_{J(t)}(\tilde{x}^s; f_i) + G_{J(t)}(\tilde{x}^s; f) \right) \\
&= G_{J(t)}(\hat{x}_t^{s+1}; f) - G_{J(t)}(\tilde{x}^s; f) + G_{J(t)}(\tilde{x}^s; f) \\
&= G_{J(t)}(\hat{x}_t^{s+1}; f)
\end{aligned} \tag{4}$$

$\hat{v}_{J(t)}^{s+1}$ is called a stochastic approximation of $G_{J(t)}(\hat{x}_t^{s+1}; f)$. More importantly, we give an upper bound for $\sum_{t=0}^{m-1} \mathbb{E} \|\hat{v}_t^{s+1}\|^2$ (Lemma 2). The lemma shows that $\hat{v}_{J(t)}^{s+1}$ would vanish after a large number of iterations. Thus, the step size γ can be set as a fixed constant, which is different to the one used in AsySZO.

As mentioned in before, $\hat{x}_t - x_t$ used in Lian et al. (2016) could not formulate the conflicts of two writing operations. In this paper, we use the following formulation to present $\hat{x}_t^{s+1} - x_t^{s+1}$.

$$x_t^{s+1} = \hat{x}_t^{s+1} - \gamma \sum_{t' \in K(t)} B_{t'}^{s+1} \hat{v}_{J(t')}^{s+1} \tag{5}$$

where $K(t)$ is a set of inner iterations, $t' \leq t-1$, $B_{t'}^{s+1}$ is a diagonal matrix with diagonal entries either 1 or 0 (0 denotes that the corresponding coordinate is overwritten by other thread). It is reasonable to assume that there exists an upper bound τ such that $\tau \geq t - \min\{t' | t' \in K(t)\}$ (i.e., Assumption 1).

Assumption 1 (Bound of delay) *There exists a upper bound τ such that $\tau \geq t - \min\{t' | t' \in K(t)\}$ for all inner iterations t in AsyDSZOVR.*

3. Convergence Analysis

In this section, we prove the convergence rate of AsyDSZOVR (Theorem 4 and Corollary 5). Specifically, we improve the convergence rate of asynchronous stochastic zeroth-order optimization from $O(\frac{1}{\sqrt{T}})$ to $O(\frac{1}{T})$. If AsyDSZOVR only uses one thread, AsyDSZOVR degenerates to the sequential doubly stochastic zeroth-order optimization algorithm with variance reduction (DSZOVR). Our theoretical analysis can work for this condition, and we have the convergence rate $\frac{1}{T}$ for DSZOVR (Corollary 6). It is also an improvement to the convergence rates of the existing sequential stochastic zeroth-order optimization algorithms (Nesterov and Spokoiny, 2011; Jamieson et al., 2012; Duchi et al., 2012; Agarwal et al., 2011).

Before providing the theoretical analysis, we give the definitions of Lipschitz constant on the original gradient, coordinated smooth function, mixtured gradient of the coordinated smooth functions, Lipschitz constant on the mixtured gradient, and the explanation of x_t^s used in the analysis as follows, which are critical to the analysis of AsyDSZOVR.

1. **Lipschitz constant on the original gradient:** For the smooth functions f_i , we have the Lipschitz constant L for ∇f_i as following.

Algorithm 2 Asynchronous Doubly Stochastic Zeroth-order Optimization with Variance Reduction (AsyDSZOVr)

Input: γ , S , and m .

Output: x^S .

```

1: Initialize  $x^0 \in \mathbb{R}^d$ ,  $p$  threads.
2: for  $s = 0, 1, 2, S - 1$  do
3:    $\tilde{x}^s \leftarrow x^s$ 
4:   All threads parallelly compute the full fake gradient  $G(\tilde{x}^s; f) = \sum_{i=1}^l \frac{1}{l} G(\tilde{x}^s; f_i)$ 
5:   For each thread, do:
6:   for  $t = 0, 1, 2, m - 1$  do
7:     Randomly sample a mini-batch  $\mathcal{B}(t)$  from  $\{1, \dots, l\}$  with equal probability.
8:     Randomly choose a set of coordinates  $J(t)$  from  $\{1, \dots, n\}$  with equal probability.
9:     Compute  $\hat{v}_{J(t)}^{s+1} = \frac{1}{|\mathcal{B}(t)|} \sum_{i \in \mathcal{B}(t)} G_{J(t)}(\hat{x}_t^{s+1}; f_i) - \frac{1}{|\mathcal{B}(t)|} \sum_{i \in \mathcal{B}(t)} G_{J(t)}(\tilde{x}^s; f_i) + G_{J(t)}(\tilde{x}^s; f)$ .
10:     $(x_{t+1}^{s+1})_{J(t)} \leftarrow \left( (x_t^{s+1}) - \gamma \hat{v}_{J(t)}^{s+1} \right)_{J(t)}$ .
11:     $(x_{t+1}^{s+1})_{\setminus J(t)} \leftarrow (x_t^{s+1})_{\setminus J(t)}$ .
12:   end for
13:    $x^{s+1} \leftarrow x_m^{s+1}$ 
14: end for

```

Assumption 2 L is the Lipschitz constant for ∇f_i ($\forall i \in \{1, \dots, l\}$) in (1). Thus, $\forall x$ and $\forall y$, L -Lipschitz smooth can be presented as

$$\|\nabla f_i(x) - \nabla f_i(y)\| \leq L\|x - y\| \quad (6)$$

Equivalently, L -Lipschitz smooth can also be written as the formulation (7).

$$f_i(x) \leq f_i(y) + \langle \nabla f_i(y), x - y \rangle + \frac{L}{2} \|x - y\|^2 \quad (7)$$

2. **Coordinated smooth function:** Given a function $f(x)$ and a predefined approximation parameter vector $[\mu_1, \mu_2, \dots, \mu_N]$, we define a coordinated smooth function $f^j(x)$ w.r.t the j -th dimension which was used in (Lian et al., 2016).

$$f^j(x) = \mathbb{E}_{v \sim U_{[-\mu_j, \mu_j]}}(p(x + ve_j)) = \frac{1}{2\mu_j} \int_{-\mu_j}^{\mu_j} f(x + ve_j) dv \quad (8)$$

where $v \sim U_{[-\mu_j, \mu_j]}$ means that v follows the uniform distribution over the interval $[-\mu_j, \mu_j]$. It should be noted that, we have the following equation between $G_j(x, f)$ and $\nabla_j f^j(x)$.

$$\begin{aligned} \nabla f^j(x) &= \frac{1}{2\mu_j} \int_{-\mu_j}^{\mu_j} \nabla_j f(x + ve_j) dv \\ &= \frac{1}{2\mu_j} (f_i(x + \mu_j e_j) - f_i(x - \mu_j e_j)) e_j = NG_j(x, f) \end{aligned} \quad (9)$$

In addition, we have

$$\mathbb{E}_j \|\nabla_j f^j(x) - \nabla_j f(x)\| \leq \frac{L^2 \sum_{j=1}^N \mu_j^2}{4N} \stackrel{\text{def}}{=} \frac{\omega}{4} \quad (10)$$

which is proved in (26) of (Lian et al., 2016).

3. **Mixed gradient of the coordinated smooth functions:** Based on the coordinated smooth function $f^j(x)$, we define a mixed gradient on the coordinated smooth functions as $\sum_{j=1}^N \nabla_j f^j(x)$.
4. **Lipschitz constant on the mixed gradient:** We assume that there exists a Lipschitz constant (\tilde{L}) on the mixed gradient as follows.

Assumption 3 \tilde{L} is the Lipschitz constant for the mixed gradient $\sum_{j=1}^N \nabla_j f^j(x)$, such that, $\forall x$ and $\forall y$, we have

$$\left\| \sum_{j=1}^N \nabla_j f^j(x) - \sum_{j=1}^N \nabla_j f^j(y) \right\| \leq \tilde{L} \|x - y\| \quad (11)$$

Because $f^j(x)$ is a smooth function of $f(x)$, it is reasonable to have a Lipschitz constant on the mixed gradient. Specifically, if $[\mu_1, \mu_2, \dots, \mu_N] = \mathbf{0}$, it is easy to verify that $\tilde{L} = L$. If $\mu_j = \infty$ for all $j = 1, \dots, N$, it is easy to verify that $\tilde{L} = 0$. Note that, it is possible that $\tilde{L} > L$.

Correspondingly, we assume there exists a relationship constant \hat{L} between the original gradient and the mixed gradient, as follows. Note that, it is also possible that $\hat{L} > 1$.

Assumption 4 For a smooth function f , we have the relationship constant \hat{L} between the original gradient and the mixed gradient as

$$\left\| \sum_{j=1}^N \nabla_j f^j(x) \right\| \leq \hat{L} \|\nabla f(x)\| \quad (12)$$

5. x_t^s : As mentioned previously, AsySBCDVR does not use any locks in the reading and writing. Thus, in the line 10 of Algorithm 2, x_t^s (left side of ' \leftarrow ') updated in the shared memory may be inconsistent with the ideal one (right side of ' \leftarrow ') computed by the proximal operator. In the analysis, we use x_t^s to denote the ideal one computed by the proximal operator. Same as mentioned in (Mania et al., 2015), there might not be an actual time the ideal ones exist in the shared memory, except the first and last iterates for each outer loop. It is noted that, x_0^s and x_m^s are exactly what is stored in shared memory. Thus, we only consider the ideal x_t^s in the analysis.

Then, we give the upper bounds of $\mathbb{E} \|G(x; f_i) - G(y; f_i)\|^2$ and $\sum_{t=0}^{m-1} \mathbb{E} \|\hat{v}_t^{s+1}\|^2$ in Lemma 1 and 2 respectively. Based on Lemma 1 and 2, we give an upper bound of $\sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2$ (Theorem 3). Then, we prove the sublinear rate of the convergence (Theorem 4 and Corollary 5).

Lemma 1 For the smooth function f_i and the corresponding approximate full gradient $G(x; f_i)$, we have

$$\mathbb{E} \|G(x; f_i) - G(y; f_i)\|^2 \leq \tilde{L}^2 \|x - y\|^2 \quad (13)$$

Proof Based on the definition of the approximate gradient $G(x; f_i)$, we have that

$$\begin{aligned} \mathbb{E} \|G(x; f_i) - G(y; f_i)\|^2 &= \mathbb{E} \left\| \frac{1}{N} \sum_{j=1}^N (G_j(x; f_i) - G_j(y; f_i)) \right\|^2 \\ &= \mathbb{E} \left\| \sum_{j=1}^N (\nabla_j f_i^j(x) - \nabla_j f_i^j(y)) \right\|^2 \leq \tilde{L}^2 \|x - y\|^2 \text{ add what is } \tilde{L} \end{aligned} \quad (14)$$

where the second equality uses (9), the first inequality uses (11). This completes the proof. \blacksquare

Lemma 2 If $Y - 2N\tilde{L}^2\gamma^2\tau^2 > 0$, we have that

$$\sum_{t=0}^{m-1} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 \leq \frac{2Y}{Y - 2N\tilde{L}^2\gamma^2\tau^2} \sum_{t=0}^{m-1} \left(\frac{2N\tilde{L}^2}{b} \|x_t^{s+1} - \tilde{x}^s\|^2 + 2\hat{L}\mathbb{E} \|\nabla f(x_t^{s+1})\|^2 \right) \quad (15)$$

Proof Let $v_t^{s+1} = \frac{1}{b} \sum_{i \in \mathcal{B}(t)} G(x_t^{s+1}; f_i) - \frac{1}{b} \sum_{i \in \mathcal{B}(t)} G(\tilde{x}^s; f_i) + G(\tilde{x}^s; f)$, we have that

$$\begin{aligned} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 &= \mathbb{E} \|\hat{v}_t^{s+1} - v_t^{s+1} + v_t^{s+1}\|^2 \\ &\leq 2\mathbb{E} \|\hat{v}_t^{s+1} - v_t^{s+1}\|^2 + 2\mathbb{E} \|v_t^{s+1}\|^2 \\ &= 2\mathbb{E} \left\| \frac{1}{b} \sum_{i \in \mathcal{B}(t)} (G(\hat{x}_t^{s+1}; f_i) - G(x_t^{s+1}; f_i)) \right\|^2 + 2\mathbb{E} \|v_t^{s+1}\|^2 \\ &\leq \frac{2}{b} \sum_{i \in \mathcal{B}(t)} \mathbb{E} \|G(\hat{x}_t^{s+1}; f_i) - G(x_t^{s+1}; f_i)\|^2 + 2\mathbb{E} \|v_t^{s+1}\|^2 \\ &\leq 2\tilde{L}^2 \mathbb{E} \|\hat{x}_t^{s+1} - x_t^{s+1}\|^2 + 2\mathbb{E} \|v_t^{s+1}\|^2 \\ &= 2\tilde{L}^2\gamma^2 \mathbb{E} \left\| \sum_{t' \in K(t)} B_{t'}^{s+1} \hat{v}_{J(t')}^{s+1} \right\|^2 + 2\mathbb{E} \|v_t^{s+1}\|^2 \\ &\leq 2\tilde{L}^2\gamma^2\tau \mathbb{E} \sum_{t' \in K(t)} \|B_{t'}^{s+1} \hat{v}_{J(t')}^{s+1}\|^2 + 2\mathbb{E} \|v_t^{s+1}\|^2 \\ &\leq 2\tilde{L}^2\gamma^2\tau \sum_{t' \in K(t)} \mathbb{E} \|\hat{v}_{J(t')}^{s+1}\|^2 + 2\mathbb{E} \|v_t^{s+1}\|^2 \\ &= \frac{2N\tilde{L}^2\gamma^2\tau}{Y} \sum_{t' \in K(t)} \mathbb{E} \|\hat{v}_{t'}^{s+1}\|^2 + 2\mathbb{E} \|v_t^{s+1}\|^2 \end{aligned} \quad (16)$$

where the first, second and fourth inequalities use the fact that $\|\sum_{i=1}^n a_i\|^2 \leq n \sum_{i=1}^n \|a_i\|^2$, the third inequality uses (13), the fifth inequality uses the Cauchy-Schwarz inequality and the fact $\|B_t^{s+1}\| \leq 1$. We consider a fixed stage $s+1$ such that $x_0^{s+1} = x_m^s$. By summing the the inequality (16) over $t = 0, \dots, m-1$, we obtain

$$\begin{aligned} \sum_{t=0}^{m-1} \mathbb{E} \|\widehat{v}_t^{s+1}\|^2 &\leq \sum_{t=0}^{m-1} \left(\frac{2N\tilde{L}^2\gamma^2\tau}{Y} \sum_{t' \in K(t)} \mathbb{E} \|\widehat{v}_{t'}^{s+1}\|^2 + 2\mathbb{E} \|v_t^{s+1}\|^2 \right) \\ &\leq \frac{2N\tilde{L}^2\gamma^2\tau^2}{Y} \sum_{t=0}^{m-1} \mathbb{E} \|\widehat{v}_t^{s+1}\|^2 + 2 \sum_{t=0}^{m-1} \mathbb{E} \|v_t^{s+1}\|^2 \end{aligned} \quad (17)$$

where the second inequality uses the Assumption 1. If $Y - 2N\tilde{L}^2\gamma^2\tau^2 > 0$, we have that

$$\sum_{t=0}^{m-1} \mathbb{E} \|\widehat{v}_t^{s+1}\|^2 \leq \frac{2Y}{Y - 2N\tilde{L}^2\gamma^2\tau^2} \sum_{t=0}^{m-1} \mathbb{E} \|v_t^{s+1}\|^2 \quad (18)$$

We next bound $\mathbb{E} \|v_t^{s+1}\|^2$ by

$$\begin{aligned} &\mathbb{E} \|v_t^{s+1}\|^2 \\ &= \mathbb{E} \left\| \frac{1}{b} \sum_{i \in \mathcal{B}(t)} G(x_t^{s+1}; f_i) - \frac{1}{b} \sum_{i \in \mathcal{B}(t)} G(\tilde{x}^s; f_i) + G(\tilde{x}^s; f) \right\|^2 \\ &= \mathbb{E} \left\| \frac{1}{b} \sum_{i \in \mathcal{B}(t)} G(x_t^{s+1}; f_i) - \frac{1}{b} \sum_{i \in \mathcal{B}(t)} G(\tilde{x}^s; f_i) + G(x^s; f) - G(x_t^{s+1}; f) + G(\tilde{x}_t^{s+1}; f) \right\|^2 \\ &\leq 2\mathbb{E} \left\| \frac{1}{b} \sum_{i \in \mathcal{B}(t)} G(x_t^{s+1}; f_i) - \frac{1}{b} \sum_{i \in \mathcal{B}(t)} G(\tilde{x}^s; f_i) - (G(x_t^{s+1}; f) - G(\tilde{x}^s; f)) \right\|^2 + 2\mathbb{E} \|G(x_t^{s+1}; f)\|^2 \\ &= \frac{2}{b^2} \mathbb{E} \left\| \sum_{i \in \mathcal{B}(t)} (G(x_t^{s+1}; f_i) - G(\tilde{x}^s; f_i) - (G(x_t^{s+1}; f) - G(\tilde{x}^s; f))) \right\|^2 + 2\mathbb{E} \left\| \frac{1}{N} \sum_{j=1}^N G_j(x_t^{s+1}; f) \right\|^2 \\ &\leq \frac{2}{b} \mathbb{E} \|G(x_t^{s+1}; f_i) - G(\tilde{x}^s; f_i) - G(x_t^{s+1}; f) + G(\tilde{x}^s; f)\|^2 + 2\mathbb{E} \left\| \frac{1}{N} \sum_{j=1}^N G_j(x_t^{s+1}; f) \right\|^2 \\ &\leq \frac{2}{b} \mathbb{E} \|G(x_t^{s+1}; f_i) - G(\tilde{x}^s; f_i)\|^2 + 2\mathbb{E} \left\| \sum_{j=1}^N \nabla_j f^j(x_t^{s+1}) \right\|^2 \\ &\leq \frac{2\tilde{L}^2}{b} \|x_t^{s+1} - \tilde{x}^s\|^2 + 2\hat{L} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 \end{aligned} \quad (19)$$

where the first inequality uses $\|\sum_{i=1}^n a_i\|^2 \leq n \sum_{i=1}^n \|a_i\|^2$, The second inequality uses Lemma 7 in (Reddi et al., 2016), the third inequality uses $\mathbb{E}\|x - \mathbb{E}x\|^2 \leq \mathbb{E}\|x\|^2$, the fourth inequality uses (13) and (12). This completes the proof. \blacksquare

Theorem 3 *Setting $c_m = 0$, $\beta_t > 0$. Let*

$$c_t = c_{t+1}(1 + \gamma\beta_t) + \left(\frac{c_{t+1}N\gamma^2}{Y} + \frac{LY\gamma^2}{2N} + \frac{\gamma^3NL^2\tau^2}{Y} \right) \frac{4YN\tilde{L}^2}{b(Y - 2N\tilde{L}^2\gamma^2\tau^2)} \quad (20)$$

$$\Gamma_t = \frac{\gamma}{2} - \left(\frac{c_{t+1}N\gamma^2}{Y} + \frac{LY\gamma^2}{2N} + \frac{\gamma^3NL^2\tau^2}{Y} \right) \frac{4Y\hat{L}}{Y - 2N\tilde{L}^2\gamma^2\tau^2} \quad (21)$$

Let η_t , β_t and c_{t+1} be chosen such that $\Gamma_t > 0$ and $\beta_t \geq 2c_{t+1}$. $\sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2$ in *AsyDSZOV*R satisfy the bound

$$\sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 \leq \frac{\mathbb{E}(f(x^s)) - \mathbb{E}(f(x^{s+1})) + \frac{\gamma N \omega m}{4}}{\min_{t \in \{0, \dots, m-1\}} \Gamma_t} \quad (22)$$

Proof We first bound $\mathbb{E} \|x_{t+1}^{s+1} - \tilde{x}^s\|^2$.

$$\begin{aligned} & \mathbb{E} \|x_{t+1}^{s+1} - \tilde{x}^s\|^2 = \mathbb{E} \|x_{t+1}^{s+1} - x_t^{s+1} + x_t^{s+1} - \tilde{x}^s\|^2 \\ &= \mathbb{E} \left(\|x_{t+1}^{s+1} - x_t^{s+1}\|^2 + \|x_t^{s+1} - \tilde{x}^s\|^2 + 2 \langle x_{t+1}^{s+1} - x_t^{s+1}, x_t^{s+1} - \tilde{x}^s \rangle \right) \\ &= \mathbb{E} \left(\gamma^2 \|\hat{v}_{J(t)}^{s+1}\|^2 + \|x_t^{s+1} - \tilde{x}^s\|^2 - 2\gamma \langle \hat{v}_{J(t)}^{s+1}, x_t^{s+1} - \tilde{x}^s \rangle \right) \\ &= \frac{N\gamma^2}{Y} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 + \mathbb{E} \|x_t^{s+1} - \tilde{x}^s\|^2 - 2\gamma \mathbb{E} \left\langle \frac{1}{b} \sum_{i \in \mathcal{B}(t)} G(\hat{x}_t^{s+1}; f_i), x_t^{s+1} - \tilde{x}^s \right\rangle \\ &\leq \frac{N\gamma^2}{Y} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 + \mathbb{E} \|x_t^{s+1} - \tilde{x}^s\|^2 + 2\gamma \mathbb{E} \left(\frac{1}{2\beta_t} \left\| \frac{1}{b} \sum_{i \in \mathcal{B}(t)} G(\hat{x}_t^{s+1}; f_i) \right\|^2 + \frac{\beta_t}{2} \|x_t^{s+1} - \tilde{x}^s\|^2 \right) \\ &= \frac{N\gamma^2}{Y} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 + (1 + \gamma\beta_t) \mathbb{E} \|x_t^{s+1} - \tilde{x}^s\|^2 + 2\gamma \mathbb{E} \left(\frac{1}{2\beta_t} \left\| \frac{1}{b} \sum_{i \in \mathcal{B}(t)} \sum_{j=1}^N \nabla_j f^j(x_t^{s+1}) \right\|^2 \right) \\ &\leq \frac{N\gamma^2}{Y} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 + (1 + \gamma\beta_t) \mathbb{E} \|x_t^{s+1} - \tilde{x}^s\|^2 + \frac{\gamma}{b\beta_t} \mathbb{E} \left(\sum_{i \in \mathcal{B}(t)} \left\| \sum_{j=1}^N \nabla_j f^j(x_t^{s+1}) \right\|^2 \right) \\ &= \frac{N\gamma^2}{Y} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 + (1 + \gamma\beta_t) \mathbb{E} \|x_t^{s+1} - \tilde{x}^s\|^2 + \frac{\gamma N}{\beta_t} \mathbb{E} \|\nabla_j f^j(x_t^{s+1})\|^2 \end{aligned} \quad (23)$$

where the first inequality uses the Young's inequality, the second inequality uses the fact that $\|\sum_{i=1}^n a_i\|^2 \leq n \sum_{i=1}^n \|a_i\|^2$. We next bound $\mathbb{E} \|\nabla_j f(x_t^{s+1}) - \nabla_j f^j(\hat{x}_t^{s+1})\|^2$.

$$\begin{aligned} & \mathbb{E} \|\nabla_j f(x_t^{s+1}) - \nabla_j f^j(\hat{x}_t^{s+1})\|^2 \\ &= \mathbb{E} \|\nabla_j f(x_t^{s+1}) - \nabla_j f(\hat{x}_t^{s+1}) + \nabla_j f(\hat{x}_t^{s+1}) - \nabla_j f^j(\hat{x}_t^{s+1})\|^2 \\ &\leq 2\mathbb{E} \|\nabla_j f(x_t^{s+1}) - \nabla_j f(\hat{x}_t^{s+1})\|^2 + 2\mathbb{E} \|\nabla_j f(\hat{x}_t^{s+1}) - \nabla_j f^j(\hat{x}_t^{s+1})\|^2 \\ &\leq \frac{2}{N} \mathbb{E} \|\nabla f(x_t^{s+1}) - \nabla f(\hat{x}_t^{s+1})\|^2 + \frac{\omega}{2} \end{aligned} \quad (24)$$

$$\begin{aligned}
&\leq \frac{2L^2}{N} \|x_t^{s+1} - \hat{x}_t^{s+1}\|^2 + \frac{\omega}{2} \\
&= \frac{2L^2\gamma^2}{N} \left\| \sum_{t' \in K(t)} B_{t'}^{s+1} \hat{v}_{J(t')}^{s+1} \right\|^2 + \frac{\omega}{2} \\
&\leq \frac{2L^2\gamma^2\tau}{N} \mathbb{E} \sum_{t' \in K(t)} \|B_{t'}^{s+1} \hat{v}_{J(t')}^{s+1}\|^2 + \frac{\omega}{2} \\
&\leq \frac{2L^2\gamma^2\tau}{N} \mathbb{E} \sum_{t' \in K(t)} \|\hat{v}_{J(t')}^{s+1}\|^2 + \frac{\omega}{2} \\
&= \frac{2L^2\gamma^2\tau}{Y} \sum_{t' \in K(t)} \mathbb{E} \|\hat{v}_{t'}^{s+1}\|^2 + \frac{\omega}{2}
\end{aligned}$$

where the first and fourth inequalities use $\|\sum_{i=1}^n a_i\|^2 \leq n \sum_{i=1}^n \|a_i\|^2$, the second inequality uses (10), the third inequality uses (6), the fifth inequality uses the Cauchy-Schwarz inequality and the fact $\|B_t^{s+1}\| \leq 1$. We bound $\mathbb{E}(f(x_{t+1}^{s+1}))$ as follows.

$$\begin{aligned}
&\mathbb{E}(f(x_{t+1}^{s+1})) \tag{25} \\
&\leq \mathbb{E} \left(f(x_t^{s+1}) + \langle \nabla f(x_t^{s+1}), x_{t+1}^{s+1} - x_t^{s+1} \rangle + \frac{L}{2} \|x_{t+1}^{s+1} - x_t^{s+1}\|^2 \right) \\
&= \mathbb{E} \left(f(x_t^{s+1}) - \gamma \langle \nabla f(x_t^{s+1}), \hat{v}_{J(t)}^{s+1} \rangle + \frac{L\gamma^2}{2} \|\hat{v}_{J(t)}^{s+1}\|^2 \right) \\
&= \mathbb{E} f(x_t^{s+1}) - \gamma \mathbb{E} \left\langle \nabla f(x_t^{s+1}), \frac{1}{b} \sum_{i \in \mathcal{B}(t)} G(\hat{x}_t^{s+1}; f_i) \right\rangle + \frac{LY\gamma^2}{2N} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 \\
&= \mathbb{E} f(x_t^{s+1}) - \gamma \mathbb{E} \left\langle \nabla f(x_t^{s+1}), \frac{1}{N} \sum_{j=1}^N G_j(\hat{x}_t^{s+1}; f) \right\rangle + \frac{LY\gamma^2}{2N} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 \\
&= \mathbb{E} f(x_t^{s+1}) - \gamma \mathbb{E} \left\langle \nabla f(x_t^{s+1}), \sum_{j=1}^N \nabla_j f^j(\hat{x}_t^{s+1}) \right\rangle + \frac{LY\gamma^2}{2N} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 \\
&= \mathbb{E} f(x_t^{s+1}) + \frac{LY\gamma^2}{2N} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 \\
&\quad - \frac{\gamma}{2} \left(\mathbb{E} \|\nabla f(x_t^{s+1})\|^2 + \mathbb{E} \left\| \sum_{j=1}^N \nabla_j f^j(\hat{x}_t^{s+1}) \right\|^2 - \mathbb{E} \left\| \nabla f(x_t^{s+1}) - \sum_{j=1}^N \nabla_j f^j(\hat{x}_t^{s+1}) \right\|^2 \right) \\
&= \mathbb{E} f(x_t^{s+1}) + \frac{LY\gamma^2}{2N} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 - \frac{\gamma}{2} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 - \frac{\gamma N}{2} \mathbb{E} \|\nabla_j f^j(\hat{x}_t^{s+1})\|^2 \\
&\quad + \frac{\gamma N}{2} \mathbb{E} \|\nabla_j f(x_t^{s+1}) - \nabla_j f^j(\hat{x}_t^{s+1})\|^2 \\
&\leq \mathbb{E} f(x_t^{s+1}) + \frac{LY\gamma^2}{2N} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 - \frac{\gamma}{2} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 - \frac{\gamma N}{2} \mathbb{E} \|\nabla_j f^j(\hat{x}_t^{s+1})\|^2
\end{aligned}$$

$$+ \frac{\gamma^3 N L^2 \tau}{Y} \sum_{t' \in K(t)} \mathbb{E} \|\hat{v}_{t'}^{s+1}\|^2 + \frac{\gamma N \omega}{4}$$

where the first inequality uses (7), the second inequality uses (24). Next, we define Lyapunov function $R_t^{s+1} = \mathbb{E} \left(f(x_t^{s+1}) + c_t \|x_t^{s+1} - \tilde{x}^s\|^2 \right)$, and give the upper bound of R_{t+1}^{s+1} as follows.

$$\begin{aligned}
& R_{t+1}^{s+1} \tag{26} \\
&= \mathbb{E} \left(f(x_{t+1}^{s+1}) + c_{t+1} \|x_{t+1}^{s+1} - \tilde{x}^s\|^2 \right) \\
&\leq \mathbb{E} f(x_t^{s+1}) + \frac{LY\gamma^2}{2N} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 - \frac{\gamma}{2} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 - \frac{\gamma N}{2} \mathbb{E} \|\nabla_j f^j(\hat{x}_t^{s+1})\|^2 \\
&\quad + \frac{\gamma^3 N L^2 \tau}{Y} \sum_{t' \in K(t)} \mathbb{E} \|\hat{v}_{t'}^{s+1}\|^2 + \frac{\gamma N \omega}{4} \\
&\quad + c_{t+1} \left(\frac{N\gamma^2}{Y} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 + (1 + \gamma\beta_t) \mathbb{E} \|x_t^{s+1} - \tilde{x}^s\|^2 + \frac{\gamma N}{\beta_t} \mathbb{E} \|\nabla_j f^j(x_t^{s+1})\|^2 \right) \\
&= \mathbb{E} f(x_t^{s+1}) + \frac{LY\gamma^2}{2N} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 - \frac{\gamma}{2} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 - \left(\frac{\gamma N}{2} - \frac{c_{t+1}\gamma N}{\beta_t} \right) \mathbb{E} \|\nabla_j f^j(\hat{x}_t^{s+1})\|^2 \\
&\quad + \frac{\gamma^3 N L^2 \tau}{Y} \sum_{t' \in K(t)} \mathbb{E} \|\hat{v}_{t'}^{s+1}\|^2 + \frac{c_{t+1}N\gamma^2}{Y} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 + c_{t+1}(1 + \gamma\beta_t) \mathbb{E} \|x_t^{s+1} - \tilde{x}^s\|^2 + \frac{\gamma N \omega}{4} \\
&\leq \mathbb{E} f(x_t^{s+1}) + \frac{LY\gamma^2}{2N} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 - \frac{\gamma}{2} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 + \frac{\gamma^3 N L^2 \tau}{Y} \sum_{t' \in K(t)} \mathbb{E} \|\hat{v}_{t'}^{s+1}\|^2 \\
&\quad + \frac{c_{t+1}N\gamma^2}{Y} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 + c_{t+1}(1 + \gamma\beta_t) \mathbb{E} \|x_t^{s+1} - \tilde{x}^s\|^2 + \frac{\gamma N \omega}{4}
\end{aligned}$$

where the first inequality uses (23) and (25), and the second inequality uses the constraint $\beta_t \geq 2c_{t+1}$. We consider a fixed stage $s+1$ such that $x_0^{s+1} = x_m^s$. By summing the the inequality (26) over $t = 0, \dots, m-1$, we obtain

$$\begin{aligned}
& \sum_{t=0}^{m-1} R_{t+1}^{s+1} \tag{27} \\
&\leq \sum_{t=0}^{m-1} \left(\mathbb{E} f(x_t^{s+1}) + \frac{LY\gamma^2}{2N} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 - \frac{\gamma}{2} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 + \frac{\gamma^3 N L^2 \tau}{Y} \sum_{t' \in K(t)} \mathbb{E} \|\hat{v}_{t'}^{s+1}\|^2 \right. \\
&\quad \left. + \frac{c_{t+1}N\gamma^2}{Y} \mathbb{E} \|\hat{v}_t^{s+1}\|^2 + c_{t+1}(1 + \gamma\beta_t) \mathbb{E} \|x_t^{s+1} - \tilde{x}^s\|^2 + \frac{\gamma N \omega}{4} \right) \\
&= \sum_{t=0}^{m-1} \left(\mathbb{E} f(x_t^{s+1}) - \frac{\gamma}{2} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 + c_{t+1}(1 + \gamma\beta_t) \mathbb{E} \|x_t^{s+1} - \tilde{x}^s\|^2 + \frac{\gamma N \omega}{4} \right. \\
&\quad \left. + \left(\frac{c_{t+1}N\gamma^2}{Y} + \frac{LY\gamma^2}{2N} + \frac{\gamma^3 N L^2 \tau}{Y} \right) \mathbb{E} \|\hat{v}_t^{s+1}\|^2 \right)
\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{t=0}^{m-1} \left(\mathbb{E}f(x_t^{s+1}) + \frac{\gamma N \omega}{4} \right. \\
&\quad \left. - \left(\frac{\gamma}{2} - \left(\frac{c_{t+1} N \gamma^2}{Y} + \frac{LY \gamma^2}{2N} + \frac{\gamma^3 N L^2 \tau^2}{Y} \right) \frac{4Y \tilde{L}}{Y - 2N \tilde{L}^2 \gamma^2 \tau^2} \right) \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 + \right. \\
&\quad \left. \left(c_{t+1}(1 + \gamma \beta_t) + \left(\frac{c_{t+1} N \gamma^2}{Y} + \frac{LY \gamma^2}{2N} + \frac{\gamma^3 N L^2 \tau^2}{Y} \right) \frac{4Y N \tilde{L}^2}{b(Y - 2N \tilde{L}^2 \gamma^2 \tau^2)} \right) \mathbb{E} \|x_t^{s+1} - \tilde{x}^s\|^2 \right) \\
&= \sum_{t=0}^{m-1} \left(R_t^{s+1} - \Gamma_t \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 + \frac{\gamma N \omega}{4} \right)
\end{aligned}$$

where the second inequality uses (15). Because $c_m = 0$, we have that $R_m^{s+1} = \mathbb{E}(f(x_m^{s+1})) = \mathbb{E}(f(x^{s+1}))$. In addition, we have that $R_0^{s+1} = \mathbb{E}(f(x_0^{s+1})) = \mathbb{E}(f(x^s))$. Based on (27), we have that

$$\begin{aligned}
\sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 &\leq \frac{\sum_{t=0}^{m-1} (R_t^{s+1} - R_{t+1}^{s+1}) + \frac{\gamma N \omega m}{4}}{\min_{t \in \{0, \dots, m-1\}} \Gamma_t} \\
&= \frac{(R_0^{s+1} - R_m^{s+1}) + \frac{\gamma N \omega m}{4}}{\min_{t \in \{0, \dots, m-1\}} \Gamma_t} \\
&= \frac{\mathbb{E}(f(x^s)) - \mathbb{E}(f(x^{s+1})) + \frac{\gamma N \omega m}{4}}{\min_{t \in \{0, \dots, m-1\}} \Gamma_t}
\end{aligned} \tag{28}$$

This completes the proof. ■

Theorem 4 Let $c_m = 0$, $\gamma = \frac{u_0 b}{L l^\alpha}$, $\beta_t = \frac{\tilde{L} N^2}{Y}$, $0 < \alpha < 1$, $0 < u_0 < 1$, and $c_t = c_{t+1}(1 + \gamma \beta_t) + \left(\frac{c_{t+1} N \gamma^2}{Y} + \frac{LY \gamma^2}{2N} + \frac{\gamma^3 N L^2 \tau^2}{Y} \right) \frac{4Y N \tilde{L}^2}{b(Y - 2N \tilde{L}^2 \gamma^2 \tau^2)}$ for $t = 0, \dots, m-1$, $b < l^\alpha$. $\frac{1}{T} \sum_{s=0}^{S-1} \sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2$ in AsyDSZOVr satisfy the bound

$$\frac{1}{T} \sum_{s=0}^{S-1} \sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 \leq \frac{\tilde{L} l^\alpha (f(x^0) - \mathbb{E}(f(x^S)))}{\sigma b T} + \frac{N u_0 \omega}{4 \sigma} \tag{29}$$

Proof Based on the specified values of γ and β_t , we have that

$$\begin{aligned}
\theta &= \gamma \beta_t + \frac{4N^2 \gamma^2 \tilde{L}^2}{b(Y - 2N \tilde{L}^2 \gamma^2 \tau^2)} = \frac{u_0 b}{\frac{Y l^\alpha}{N^2}} + \frac{4u_0^2 b}{\frac{Y l^{2\alpha}}{N^2} - \frac{2\tau^2 u_0^2 b^2}{N}} \\
&= \frac{u_0 b N^2}{Y l^\alpha} + \frac{4u_0^2 b N^2}{Y l^{2\alpha} - 2N \tau^2 u_0^2 b^2} \\
&\leq \frac{5u_0 b N^2}{Y l^\alpha}
\end{aligned} \tag{30}$$

where the inequality uses the constraint $Y l^\alpha \leq Y l^{2\alpha} - 2N \tau^2 u_0^2 b^2$ by appropriately choosing α and u_0 . We set $m = \lfloor \frac{Y l^\alpha}{5u_0 b N^2} \rfloor$, from the recurrence definition of c_t , we have that

$$c_0 = \frac{4Y N \tilde{L}^2}{b(Y - 2N \tilde{L}^2 \gamma^2 \tau^2)} \left(\frac{LY \gamma^2}{2N} + \frac{\gamma^3 N L^2 \tau^2}{Y} \right) \frac{(1 + \theta)^m - 1}{\theta} \tag{31}$$

$$\begin{aligned}
&= \frac{4YN\tilde{L}^2}{b(Y - 2N\tilde{L}^2\gamma^2\tau^2)} \frac{\frac{LYu_0^2b^2}{2NL^2l^{2\alpha}} + \frac{NL^2\tau^2u_0^3b^3}{YL^3l^{3\alpha}}}{\frac{u_0bN^2}{Yl^\alpha} + \frac{4u_0^2bN^2}{Yl^{2\alpha} - 2N\tau^2u_0^2b^2}} ((1 + \theta)^m - 1) \\
&\leq \frac{4YN\tilde{L}^2l^{2\alpha}}{b(Yl^{2\alpha} - 2N\tau^2u_0^2b^2)} \frac{\frac{LYu_0^2b^2}{2N} + \frac{NL^2\tau^2u_0^3b^3}{Y}}{5u_0^2bN^2\tilde{L}^2l^{2\alpha}} ((1 + \theta)^m - 1) \\
&= \frac{\frac{2LY^2}{N} + 4NL^2\tau^2u_0b}{5N} ((1 + \theta)^m - 1) \\
&\leq \underbrace{\frac{\frac{2LY^2}{N} + 4NL^2\tau^2u_0b}{5N}}_{:=\varrho_1} (e - 1)
\end{aligned}$$

where the first inequality uses $\tilde{L}^3l^{3\alpha} \geq \tilde{L}^2l^{2\alpha}$, the second inequality uses the fact $(1 + \frac{1}{a})^a$ is increasing for $a > 0$, and $\lim_{a \rightarrow \infty} (1 + \frac{1}{a})^a = e$, which is also used in (Reddi et al., 2015). Let $\tilde{\Gamma}$ denote the following quantity:

$$\tilde{\Gamma} = \min_{t \in \{0, \dots, m-1\}} \frac{\gamma}{2} - \left(\frac{c_{t+1}N\gamma^2}{Y} + \frac{LY\gamma^2}{2N} + \frac{\gamma^3NL^2\tau^2}{Y} \right) \frac{4Y\hat{L}}{Y - 2N\tilde{L}^2\gamma^2\tau^2} \quad (32)$$

Now we give a lower bound of $\tilde{\Gamma}$ as

$$\begin{aligned}
\tilde{\Gamma} &= \min_{t \in \{0, \dots, m-1\}} \frac{\gamma}{2} - \left(\frac{c_{t+1}N\gamma^2}{Y} + \frac{LY\gamma^2}{2N} + \frac{\gamma^3NL^2\tau^2}{Y} \right) \frac{4Y\hat{L}}{Y - 2N\tilde{L}^2\gamma^2\tau^2} \quad (33) \\
&\geq \frac{\gamma}{2} - \left(\frac{c_0N\gamma^2}{Y} + \frac{LY\gamma^2}{2N} + \frac{\gamma^3NL^2\tau^2}{Y} \right) \underbrace{\frac{4Y\hat{L}}{Y - 2N\tilde{L}^2\gamma^2\tau^2}}_{:=\varrho_2} \\
&= \frac{\gamma}{2} - \left(\frac{\varrho_1N\gamma^2}{Y} + \frac{LY\gamma^2}{2N} + \frac{\gamma^3NL^2\tau^2}{Y} \right) \varrho_2 \\
&\geq \underbrace{\gamma \left(\frac{1}{2} - \frac{\varrho_1N\varrho_2\gamma}{Y} - \frac{LY\varrho_2\gamma}{2N} - \frac{\varrho_2NL^2\tau^2\gamma^2}{Y} \right)}_{\varrho_3} \\
&\geq \frac{\sigma b}{\tilde{L}l^\alpha}
\end{aligned}$$

where the first inequality holds because c_t decrease with t , ϱ_2 are constants, $\sigma = \varrho_3u_0$. For the last inequality, we use the constraint $b < l^\alpha$. Thus, we can appropriately choose a value of u_0 , such that $\varrho_3 > 0$, and σ is a small value independent to l .

$$\begin{aligned}
\frac{1}{T} \sum_{s=0}^{S-1} \sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 &\leq \frac{1}{T} \sum_{s=0}^{S-1} \frac{\mathbb{E}(f(x^s)) - \mathbb{E}(f(x^{s+1})) + \frac{\gamma N \omega m}{4}}{\tilde{\Gamma}} \quad (34) \\
&= \frac{f(x^0) - \mathbb{E}(f(x^S)) + \frac{\gamma N \omega T}{4}}{T\tilde{\Gamma}} \\
&\leq \frac{\tilde{L}l^\alpha (f(x^0) - \mathbb{E}(f(x^S)))}{\sigma b T} + \frac{Nu_0\omega}{4\sigma}
\end{aligned}$$

This completes the proof. ■

Corollary 5 Let $c_m = 0$, $\gamma = \frac{u_0 b}{L l^\alpha}$, $\beta_t = \frac{\tilde{L} N^2}{Y}$, $0 < \alpha < 1$, $0 < u_0 < 1$, and $c_t = c_{t+1}(1 + \gamma\beta_t) + \left(\frac{c_{t+1} N \gamma^2}{Y} + \frac{L Y \gamma^2}{2N} + \frac{\gamma^3 N L^2 \tau^2}{Y}\right) \frac{4 Y N \tilde{L}^2}{b(Y - 2 N \tilde{L}^2 \gamma^2 \tau^2)}$ for $t = 0, \dots, m-1$, $b < l^\alpha$. If $\omega = 0$, $\frac{1}{T} \sum_{s=0}^{S-1} \sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2$ in AsyDSZOVR satisfy the bound

$$\frac{1}{T} \sum_{s=0}^{S-1} \sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 \leq \frac{\tilde{L} l^\alpha (f(x^0)) - \mathbb{E}(f(x^S))}{\sigma b T} \quad (35)$$

Corollary 6 Let $c_m = 0$, $\gamma = \frac{u_0 b}{L l^\alpha}$, $\beta_t = \frac{\tilde{L} N^2}{Y}$, $0 < \alpha < 1$, $0 < u_0 < 1$, and $c_t = c_{t+1}(1 + \gamma\beta_t) + \left(\frac{c_{t+1} N \gamma^2}{Y} + \frac{L Y \gamma^2}{2N} + \frac{\gamma^3 N L^2 \tau^2}{Y}\right) \frac{4 Y N \tilde{L}^2}{b(Y - 2 N \tilde{L}^2 \gamma^2 \tau^2)}$ for $t = 0, \dots, m-1$, $b < l^\alpha$. $\frac{1}{T} \sum_{s=0}^{S-1} \sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2$ in DSZOVR satisfy the bound

$$\frac{1}{T} \sum_{s=0}^{S-1} \sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 \leq \frac{\tilde{L} l^\alpha (f(x^0)) - \mathbb{E}(f(x^S))}{\sigma b T} + \frac{N u_0 \omega}{4\sigma} \quad (36)$$

If $\omega = 0$, $\frac{1}{T} \sum_{s=0}^{S-1} \sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2$ in DSZOVR satisfy the bound

$$\frac{1}{T} \sum_{s=0}^{S-1} \sum_{t=0}^{m-1} \mathbb{E} \|\nabla f(x_t^{s+1})\|^2 \leq \frac{\tilde{L} l^\alpha (f(x^0)) - \mathbb{E}(f(x^S))}{\sigma b T} \quad (37)$$

4. Conclusion

In this paper, we propose an asynchronous doubly stochastic zeroth-order optimization algorithm using the accelerated technology of variance reduction (AsyDSZOVR). Our AsyDSZOVR randomly select a set of samples and a set of features simultaneously to handle large scale problems both in volume and dimension. Rigorous theoretical analysis show that the convergence rate can be improved from $O(\frac{1}{\sqrt{T}})$ the best result of existing algorithms to $O(\frac{1}{T})$. Also our theoretical results is an improvement to the ones of the sequential stochastic zeroth-order optimization algorithms.

References

- Alekh Agarwal, Dean P Foster, Daniel J Hsu, Sham M Kakade, and Alexander Rakhlin. Stochastic convex optimization with bandit feedback. In *Advances in Neural Information Processing Systems*, pages 1035–1043, 2011.
- Sébastien Bubeck and Nicolo Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *arXiv preprint arXiv:1204.5721*, 2012.
- Richard H Byrd, SL Hansen, Jorge Nocedal, and Yoram Singer. A stochastic quasi-newton method for large-scale optimization. *SIAM Journal on Optimization*, 26(2):1008–1031, 2016.

- John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.
- David A Freedman. *Statistical models: theory and practice*. cambridge university press, 2009.
- Zhouyuan Huo and Heng Huang. Asynchronous stochastic gradient descent with variance reduction for non-convex optimization. *arXiv preprint arXiv:1604.03584*, 2016.
- Kevin G Jamieson, Robert Nowak, and Ben Recht. Query complexity of derivative-free optimization. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2012.
- Xiangru Lian, Yijun Huang, Yuncheng Li, and Ji Liu. Asynchronous parallel stochastic gradient for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 2737–2745, 2015.
- Xiangru Lian, Huan Zhang, Cho-Jui Hsieh, Yijun Huang, and Ji Liu. A comprehensive linear speedup analysis for asynchronous stochastic parallel optimization from zeroth-order to first-order. *arXiv preprint arXiv:1606.00498*, 2016.
- Ji Liu and Stephen J Wright. Asynchronous stochastic coordinate descent: Parallelism and convergence properties. *SIAM Journal on Optimization*, 25(1):351–376, 2015.
- Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. *arXiv preprint arXiv:1507.06970*, 2015.
- Yurii Nesterov and Vladimir Spokoiny. Random gradient-free minimization of convex functions. *Foundations of Computational Mathematics*, pages 1–40, 2011.
- Meisam Razaviyayn, Mingyi Hong, Zhi-Quan Luo, and Jong-Shi Pang. Parallel successive convex approximation for nonsmooth nonconvex optimization. In *NIPS*, 2014.
- Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 693–701, 2011.
- Sashank J Reddi, Ahmed Hefny, Suvrit Sra, Barnabas Poczos, and Alex J Smola. On variance reduction in stochastic gradient descent and its asynchronous variants. In *Advances in Neural Information Processing Systems*, pages 2647–2655, 2015.
- Sashank J Reddi, Suvrit Sra, Barnabas Poczos, and Alex Smola. Fast stochastic methods for nonsmooth nonconvex optimization. *arXiv preprint arXiv:1605.06900*, 2016.
- Xia Shen, Moudud Alam, Freddy Fikse, and Lars Rönnegård. A novel generalized ridge regression method for quantitative genetics. *Genetics*, 193(4):1255–1268, 2013.
- Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

- Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd international conference on Machine learning*, pages 896–903. ACM, 2005.
- Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.
- Yang You, Xiangru Lian, Ji Liu, Hsiang-Fu Yu, Inderjit S Dhillon, James Demmel, and Cho-Jui Hsieh. Asynchronous parallel greedy coordinate descent. In *Advances In Neural Information Processing Systems*, pages 4682–4690, 2016.
- Shen-Yi Zhao and Wu-Jun Li. Fast asynchronous parallel stochastic gradient descent: A lock-free approach with convergence guarantee. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.